DOCUMENT RESUME

ED 171 793 TM 009 436

AUTHOR Kleinke, David J.
TITLE Item Order, Response Format, and Examinee Sex and Handedness and Performance on a Multiple-Choice Test.
PUB DATE Apr 79
NOTE 16p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; *Answer Sheets; *Complexity Level; Eye Hand Coordination; Grade 4; Intermediate Grades; Lateral Dominance; *Left Handed Writer; Multiple Choice Tests; Response Mode; Response Style (Tests); *Sex Differences; *Test Construction; Test Items; *Timed Tests
IDENTIFIERS *Test Format

ABSTRACT

Four forms of a 36-item adaptation of the Stanford Achievement Test were administered to 484 fourth graders. External factors potentially influencing test performance were examined, namely: (1) item order (easy-to-difficult vs. uniform); (2) response location (left column vs. right column); (3) handedness which may interact with response location; and (4) sex. The total test score, a speed score, and a speeded items-correct score were the dependent variables. Three univariate 2x4 analyses of variance revealed no significant interaction effects. However, the main effect of sex was significant on all three analyses, boys outperforming girls. On the two speed-related scores, left-handed students and students taking easy-to-difficult forms were superior. The last finding is consistent with previous studies, but the sex and handedness findings were unexpected. If a test is not timed, item order and response location apparently make no difference. In conclusion, teachers constructing power tests should consider factors other than sex or handedness when deciding on order or format. (Author/CP)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED171793

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

# Item Order, Response Format, and Examinee Sex and Handedness and Performance on a Multiple-Choice Test

David J. Kleinke

Syracuse University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

David J. Kleinke

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Presented to annual meeting of National Council on Measurement in Education. San Francisco; April, 1979.

TM009 436

## Abstract

Four forms of a 36-item test composed of Stanford Achievement Test items were administered to 484 fourth graders. Two item orders were used: easy-to-difficult and uniform. Two in-booklet response locations were also employed: left- and right-column. The total-test score, a speed score, and a speeded items-correct score were the dependent variables. Three univariate $2^4$ analyses of variance revealed no significant interaction effects. The main effect of sex was significant on all three, boys outperforming girls. On the two speed-related scores, left-handed pupils and examinees taking easy-to-difficult forms were superior. The last finding is consistent with previous studies, which used college students as subjects. It is concluded that teachers designing power tests use considerations other than pupils' sex or handedness when deciding on order or format.

# Item Order, Response Format, and Examinee Sex and Handedness and Performance on a Multiple-Choice Test

An examinee's score on an achievement test ought to be solely a function of the examinee's mastery of the subject being tested. External factors should have no influence on test performance. The study being reported here is one of the effect of some such external influences, alone and in combination.

The external factors investigated were (a) the order in which the items are arranged in the test booklet, (b) the location, in the test booklet, of the response space, (c) the hand the examinee uses to respond to the items, and (d) the examinee's sex. Each of these might influence test performance alone. Each could function in combination with one or more of the others to affect test performance.

## Background

### Item Order

Measurement textbooks often suggest that test items be arranged from easy to difficult within a test. For examples, see Ahmann and Glock (1971), Sax (1974), and Stanley and Hopkins (1972). Beginning with easy items is alleged to be beneficial to the examinee in two ways. First, the examinee is positively reinforced in some way by being able to answer items easily. This would lead to a continuation of purposeful test-taking behaviors. Second, the examinee is not negatively reinforced in the early stages of the test. Few or no adverse stimuli, in the form of difficult or confusing items, are

presented. Thus, one would expect examinees to obtain higher scores on tests whose items are arranged from easy to difficult than on tests with other item arrangements.

Empirical studies have not, in general, supported this expectation. (For examples, see Marso, 1970; Munz & Smouse, 1968; Sax & Cromack, 1966; Towle & Merrill, 1975.) In the typical study, at least two versions of the same test are employed. One of the versions has the items arranged from easy to difficult. Others are arranged randomly, uniformly, or even from difficult to easy. The findings generally have been that the mean of the easy-to-difficult form is not different from the mean(s) of the other form(s). The only exception to this general finding is for tests with stringent time limits. For instance, the examinees in Sax and Cromack's easy-to-difficult group obtained higher scores than the others when the test was composed of 70 quantitative items with a 30-minute time limit.

A serious limitation of these studies has been their use of either undergraduates or graduate students as examinees. It is reasonable to expect that these students, as a group, are more intelligent, more persistent, and more sophisticated test-takers than are examinees in the schools.

Response Location

Many objective tests provide for responses to be made in the test booklet. This is especially true of locally-constructed achievement tests for elementary-school pupils. For ease of scoring, the response spaces are often arranged in columns to either the left of the items

or the right of the items. There is no agreement among measurement specialists as to which is the preferred side. For instance, Ahmann and Glock (1971, pp. 91-92) show examples of left-column answers, while Thorndike (1971, p.173) provides spaces at the right.

One cannot hypothesize that either the left-column placement or the right-column placement would lead to superior test performance by itself. However, this aspect of test format could interact with pupil handedness to lead to higher (or lower) scores.

Handedness

The problem is that some pupils must cover a question with their writing hand in order to make a response. This is especially true of many left-handed pupils responding to "right-handed" questions.

It was not a purpose of this study to investigate the effects of handedness per se. No overall mean differences are to be expected when comparing left-handed children and right-handed children (Hardyck, Petrinovich, & Goldman, 1976; Keller, Croake, & Riesenman, 1973). The concern in this study was the interaction between handedness and test format.

Sex

Similarly, pupil gender was included as it might contribute to an interaction, not as a main effect of interest. It was hypothesized that there might be a three-way interaction involving sex, handedness, and test format. Left-handed boys could be penalized when given right-handed tests, but left-handed girls might not be subject to the same problem.

## Method

### Instrument and Subjects

The test consisted of the first 36 items of the Social Sciences Test of the 1973 Stanford Achievement Test, Intermediate I (Madden, Gardner, Rudman, Karlsen, & Merwin, 1973)[1]. Seven items were modified slightly for editorial purposes. There were two item orders and two response formats, so there were four forms.

The items in two of the forms (LE and RE) were arranged from easy to difficult. The item difficulty values used for this were the mid-values of the difficulty indices reported in the Stanford technical manual. These item difficulty values were also used to assemble the "uniform" forms, LU and RU. For these forms, the items were divided into six blocks of six items each. The first block contained the six most difficult items; the next block contained the six next most difficult; and so forth. The item order for forms LU and RU was determined by randomly selecting one item to be included in each set of six item numbers from each of the six difficulty blocks. Thus, items 1 through 6 contained one of the six most difficult items, one of the six next most difficult items, and so forth, as did items 7 through 12, 13 through 18,...and 31 through 36.

The response locations also varied. Pupils responded to forms RE and RU at the right of the page. The response spaces were in the left-hand columns of forms LE and LU.

The test was administered during May 1978 to the 486 fourth-grade pupils in a predominately middle-class suburban school district[2] who

were present on the day of testing. Two pupils' papers could not be used. The distribution of examinees by sex, handedness, and test form is presented in Table 1.

---

Insert Table 1 about here

---

Procedures

The tests were administered in the pupils' regular classrooms by their teachers. The teachers also recorded each pupil's sex and handedness on the cover page of the pupil's examination booklet. The in-booklet responses were subsequently transferred to standard answer sheets for machine processing. Pupils' names were not coded on the answer sheets. As a condition of permission to do this study, individual examinees were not to be identified.

Twenty minutes were allowed for the actual testing. At the end of the first ten minutes, the teachers announced that the pupils should draw a line "under the question you are now reading" and then continue with the test.

Data Analysis

There were thus three scores for each pupil. The first score, "TOTAL," was the number of items answered correctly on the entire test. The second score was designated "SPEED." This was one less than the number of the item under which the line was drawn. For example, if the line was drawn under item 27, it was assumed that the pupil had completed 26 questions, for a SPEED score of 26. "TOT10" was the score the pupil was assumed to have obtained in the

first ten minutes. If a pupil had a speed score of 26, TOT10 is the number of items correct from only items 1 through 26.

Initial scoring and data analysis were performed using a locally-developed item- and test analysis package. Subsequent analyses were performed using SPSS (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975). The central analyses were three $2^4$ analyses of variance, using the SPSS ANOVA subprogram with defaults. The defaults are to examine all interaction effects and to adjust each independent variable main effect for the other main effects (Nie et al., 1975, p. 418).

## Results

First, the item ordering with respect to difficulty appears to have been proper. The Spearman rank correlation between item number and actual difficulty was -.90 for forms LE and RE combined. For forms LU and RU combined, the correlation was .03.

None of the interactions in any of the analyses of variance resulted in a statistically significant *F* ratio. For clarity, then, the means and standard deviations associated with the main effects are presented in Table 2.

---

Insert Table 2 about here

---

## Discussion

The effect of item order was consistent with earlier findings (Sax & Cromack, 1966, Towle & Merrill, 1975). If examinees have ample time to complete the test, then the item order with respect to difficulty

does not affect performance. If the test is speeded, then the easy-to-difficult ordering produces higher scores. In this respect at least, fourth graders perform the same way students of educational psychology or educational measurement do.

The expected effect of response format either alone or, especially, interacting with examinee handedness did not occur. If there is such a phenomenom, it is probably overcome by late in the fourth grade.

The statistically significant differences associated with both sex and handedness as main effects were unexpected. No differences were expected as a function of either sex or handedness (Hardyck, et al., 1976; Keller, et al., 1973).

There was considerable loss of examinees for the analyses involving SPEED and TOT10. A total of 170 pupils failed to draw the requested line after ten minutes. Before proceeding with the SPEED and TOT10 analyses, several checks were made to determine whether there were important differences between the 314 included in these analyses and those who were excluded. There were no such differences, using an alpha level of .25 to avoid Type II errors. The 170 were distributed among all classrooms, indicating that every teacher read the required directions. Sex, handedness, and booklet assignments were essentially identical. Finally, performance on TOTAL was not significantly different for the two groups.

The choice of univariate analyses of variance, as opposed to a multivariate analysis of variance, was based on several considerations.

Being able to combine speed and accuracy scores into TOT10 greatly improves interpretability over having them forced into a best linear combination. Also, if significant multivariate Fs are encountered, one then proceeds to perform and interpret the univariate Fs to understand "what really happened." Given that there were only two levels of each of the four independent variables this would have produced redundant computations.

## Conclusions

Except when a test is speeded, it apparently makes no difference what the item order is or where the in-booklet responses are to be recorded. This is apparently true for end-of-year fourth-grade pupils, regardless of their sex or handedness. Whether this conclusion is applicable to younger pupils is still open to investigation.

In constructing a power test of academic achievement, a teacher might as well order the items by difficulty. The current study, as well as all of its predecessors, failed to uncover any harmful effects of ordering by difficulty. At least this order ought to randomize item content. Randomizing item content will reduce the possibility that an examinee will use the location of an item as a clue to its answer. As for choosing the test format, the teacher is probably best advised to use the location that better facilitates scoring.

## References

Ahmann, J. S. & Glock, M. D. Measuring and evaluating educational achievement. Boston: Allyn and Bacon, 1971.

Hardyck, C., Petrinovich, L. F. & Goldman, R. D. Left-handedness and cognitive deficit. Cortex, 1976, 12, 266-279.

Keller, J. F., Croake, J. W., & Riesenman, C. Relationships among handedness, intelligence, sex, and reading achievement of school age children. Perceptual & Motor Skills, 1973, 37, 159-162.

Madden, R., Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. Stanford achievement tests. New York: Harcourt Bruce Javanovich (now, the Psychological Corporation), 1973.

Marso, R. N. Test item arrangement, testing time, and performance. Journal of Educational Measurement, 1970, 7, 113-118.

Munz, D. C. & Smouse, A. D. Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. Journal of Educational Psychology, 1968, 59, 370-374.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1975.

Sax, G. Principles of educational measurement and evaluation. Belmont, CA: Wadsworth, 1974.

Sax, G. & Cromack, T. R. The effects of various forms of item arrangements on test performances. Journal of Educational Measurement, 1966, 3, 309-311.

Stanley, J. C. & Hopkins, K. D. Educational and psychological measurement and evaluation. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Thorndike, R. L. Reproducing the test. In R. L. Thorndike (Ed.) Educational measurement (2nd ed.). Washington: American Council on Education, 1971.

Towle, N. J. & Merrill, P. F. Effects of anxiety type and item-difficulty sequencing on mathematics test performance. Journal of Educational Measurement, 1975, 12, 241-249.

## Footnotes

[1]Thanks are due to the authors and publisher for permission to revise and reproduce the test.

[2]The cooperation of the administration and faculty of the West Genesee Central School District, Camillus, NY is gratefully acknowledged.

Table 1

Numbers of Examinees, by Handedness, Sex, and Test Form

| Form | Handedness | | | | Total |
|---|---|---|---|---|---|
| | Left | | Right | | |
| | Girls | Boys | Girls | Boys | |
| LE | 5 | 8 | 63 | 47 | 123 |
| RE | 8 | 6 | 59 | 47 | 120 |
| LU | 7 | 6 | 55 | 53 | 121 |
| RU | 7 | 7 | 51 | 55 | 120 |
| Total | 27 | 27 | 228 | 202 | 484 |

Table 2

Numbers of Examinees, Means, and Standard Deviation for Scores, by Independent Variable

| Independent Variable | Score | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | TOTAL | | | SPEED | | | TOT10 | | |
| | N | Mean | SD | N | Mean | SD | N | Mean | SD |
| Order | | | | | | | | | |
| Easy-to-difficult | 243 | 22.5 | 5.4 | 146 | 29.7[b] | 5.0 | 146 | 19.5[b] | 5.6 |
| Uniform | 241 | 21.8 | 5.6 | 168 | 27.8 | 6.1 | 168 | 16.7 | 6.5 |
| Format | | | | | | | | | |
| Left column | 244 | 22.2 | 5.7 | 154 | 28.1 | 5.7 | 154 | 17.6 | 6.4 |
| Right column | 240 | 22.1 | 5.3 | 160 | 29.2 | 5.6 | 160 | 18.4 | 6.1 |
| Sex | | | | | | | | | |
| Boys | 229 | 23.1[c] | 5.2 | 153 | 30.0[c] | 5.3 | 153 | 19.8[c] | 6.1 |
| Girls | 255 | 21.2 | 5.6 | 161 | 27.4 | 5.8 | 161 | 16.3 | 5.9 |
| Handedness | | | | | | | | | |
| Left-handed | 54 | 22.9 | 5.4 | 39 | 31.2[b] | 4.3 | 39 | 20.0[a] | 6.3 |
| Right-handed | 430 | 22.0 | 5.5 | 275 | 28.3 | 5.8 | 275 | 17.7 | 6.2 |
| Total | 484 | 22.1 | 5.5 | 314 | 28.7 | 5.7 | 314 | 18.0 | 6.3 |

[a] $p < .05$, two-tailed

[b] $p < .01$, two-tailed

[c] $p < .001$, two-tailed